Towards Expert-Level Medical Question Answering with Large Language Models

Year: 2023 | Citations: 663 | Authors: K. Singhal, Tao Tu, Juraj Gottweis, R. Sayres, Ellery Wulczyn

---

Abstract

Recent artificial intelligence (AI) systems have reached milestones in"grand challenges"ranging from Go to protein-folding. The capability to retrieve medical knowledge, reason over it, and answer medical questions comparably to physicians has long been viewed as one such grand challenge. Large language models (LLMs) have catalyzed significant progress in medical question answering; Med-PaLM was the first model to exceed a"passing"score in US Medical Licensing Examination (USMLE) style questions with a score of 67.2% on the MedQA dataset. However, this and other prior work suggested significant room for improvement, especially when models' answers were compared to clinicians' answers. Here we present Med-PaLM 2, which bridges these gaps by leveraging a combination of base LLM improvements (PaLM 2), medical domain finetuning, and prompting strategies including a novel ensemble refinement approach. Med-PaLM 2 scored up to 86.5% on the MedQA dataset, improving upon Med-PaLM by over 19% and setting a new state-of-the-art. We also observed performance approaching or exceeding state-of-the-art across MedMCQA, PubMedQA, and MMLU clinical topics datasets. We performed detailed human evaluations on long-form questions along multiple axes relevant to clinical applications. In pairwise comparative ranking of 1066 consumer medical questions, physicians preferred Med-PaLM 2 answers to those produced by physicians on eight of nine axes pertaining to clinical utility ($p<0.001$). We also observed significant improvements compared to Med-PaLM on every evaluation axis ($p<0.001$) on newly introduced datasets of 240 long-form"adversarial"questions to probe LLM limitations. While further studies are necessary to validate the efficacy of these models in real-world settings, these results highlight rapid progress towards physician-level performance in medical question answering.